

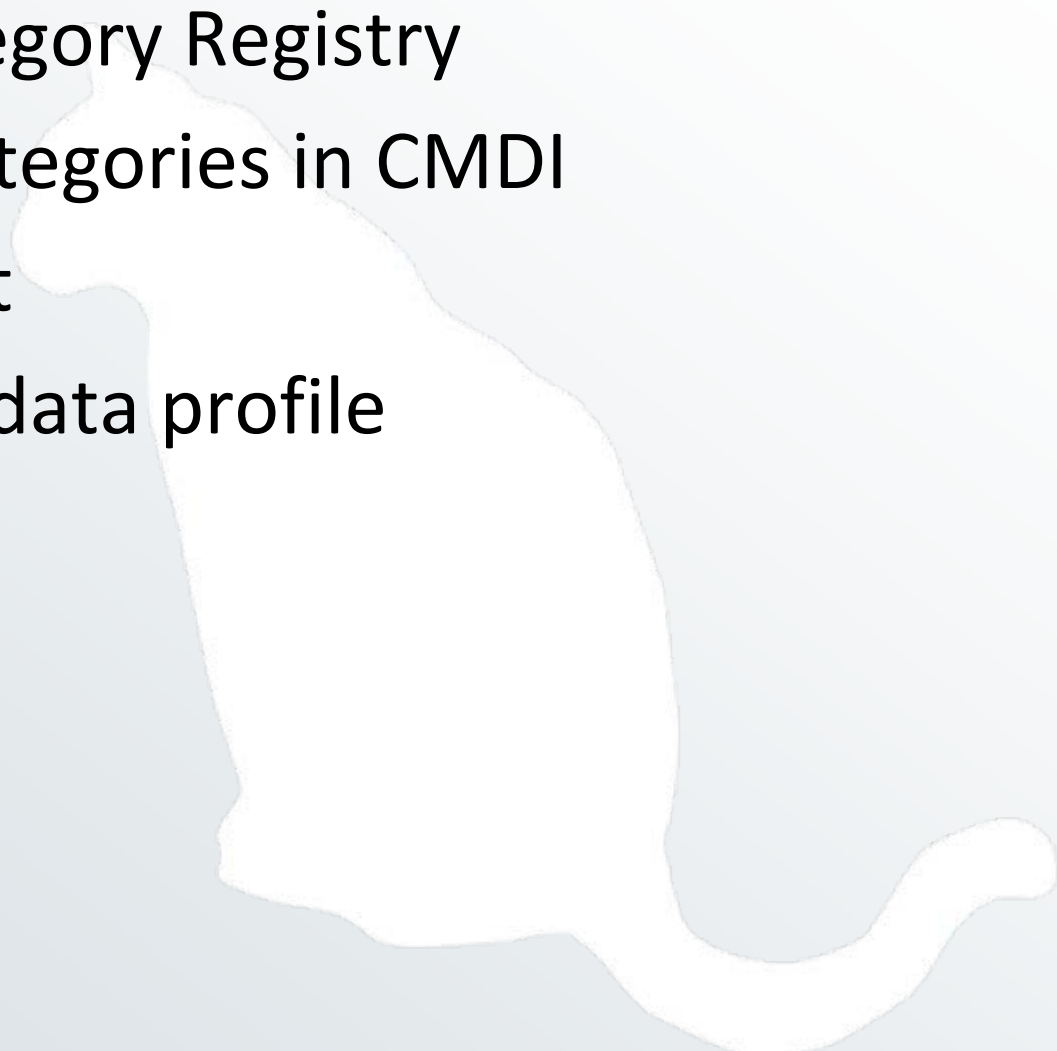
Use of ISOcat within CMDI

Menzo Windhouwer

Marc Kemps-Snijders

Sue Ellen Wright

Outline

- ISOcat: a Data Category Registry
 - The role of data categories in CMDI
 - A glimpse of ISOcat
 - Status of the metadata profile
 - A matter of trust
 - Upcoming
- 

ISOcat: a Data Category Registry

- The reference implementation of ISO 12620:2009
 - Terminology and other content and language resources — Specification of data categories and management of a Data Category Registry for language resources
- A data category
 - is the result of the specification of a given data field
 - an elementary descriptor in a linguistic structure or an annotation scheme

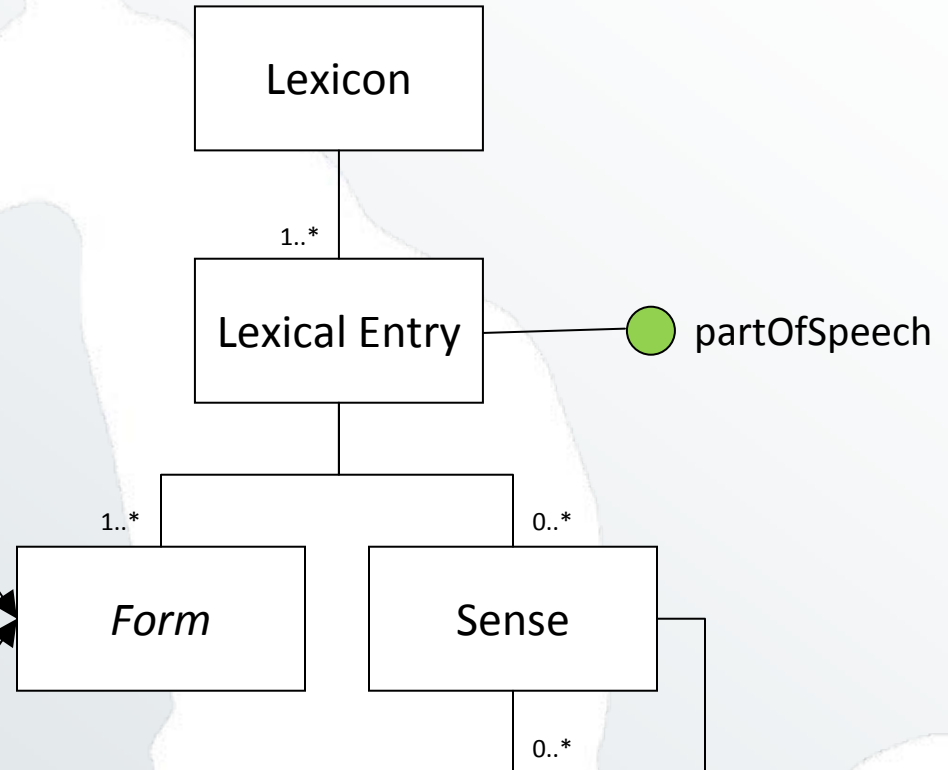
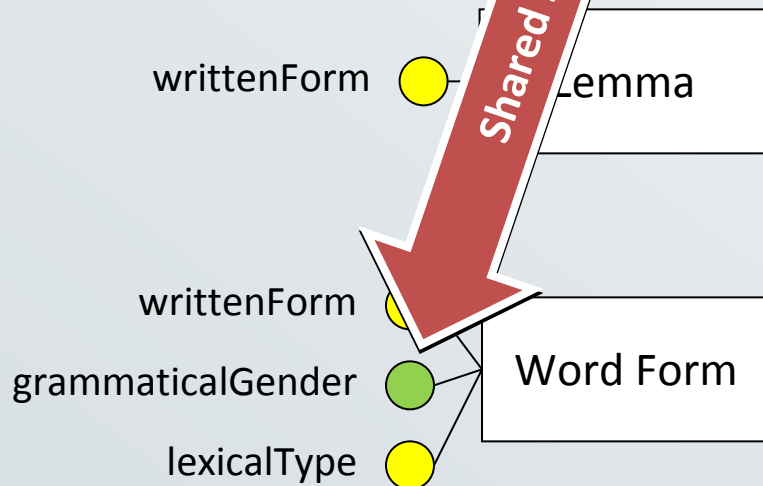
Data categories and linguistic resources

Language	BWO	...

wordOrder ●

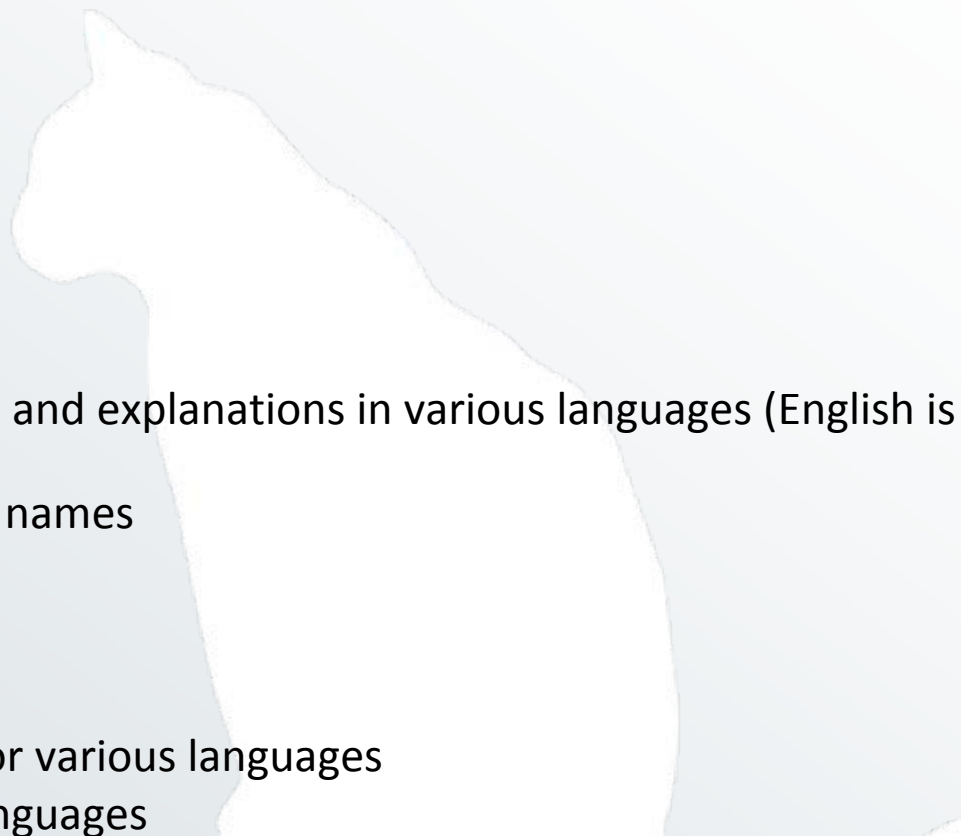
grammaticalGender ●

A (schema for a) typological database



A (schema for a) lexicon

Data category specification

- Administrative information
 - Identifier
 - Version
 - Origin
 - Justification
 - Status
 - Descriptive information
 - Names, definitions, examples and explanations in various languages (English is mandatory)
 - Application (domain) specific names
 - Conceptual domain
 - Possible values (per profile)
 - Linguistic information
 - Examples and explanations for various languages
 - Possible values for various languages
- 

The role of data categories in CMDI

- CMD components, elements and items can have links to *concepts*
- These links should be *resolvable* to a concept description
- This concept description gives *explicit semantics*
- Elements and components can use different terminology but still have *common semantics*

- ISOcat provides resolvable links to the semantic description of data categories (DCs)
 - CMD items: simple DCs
 - CMD elements: complex DCs
 - CMD components: container DCs (upcoming)

Data category references in CMDI

```
<CMD_Component name="HeadWordType">
  <CMD_Element name="HeadWordType"
    ConceptLink="http://www.isocat.org/datcat/DC-2486">
    <ValueScheme>
      <enumeration>
        <item ConceptLink="http://www.isocat.org/datcat/DC-286">Lemma</item>
        <item ConceptLink="http://www.isocat.org/datcat/DC-2948">Word form</item>
        <item ConceptLink="http://www.isocat.org/datcat/DC-350">Phrase</item>
        <item ConceptLink="http://www.isocat.org/datcat/DC-1386">Sentence</item>
        <item ConceptLink="http://www.isocat.org/datcat/DC-2599">Other</item>
        <item ConceptLink="http://www.isocat.org/datcat/DC-2592">Unspecified</item>
      </enumeration>
    </ValueScheme>
  </CMD_Element>
</CMD_Component>
```

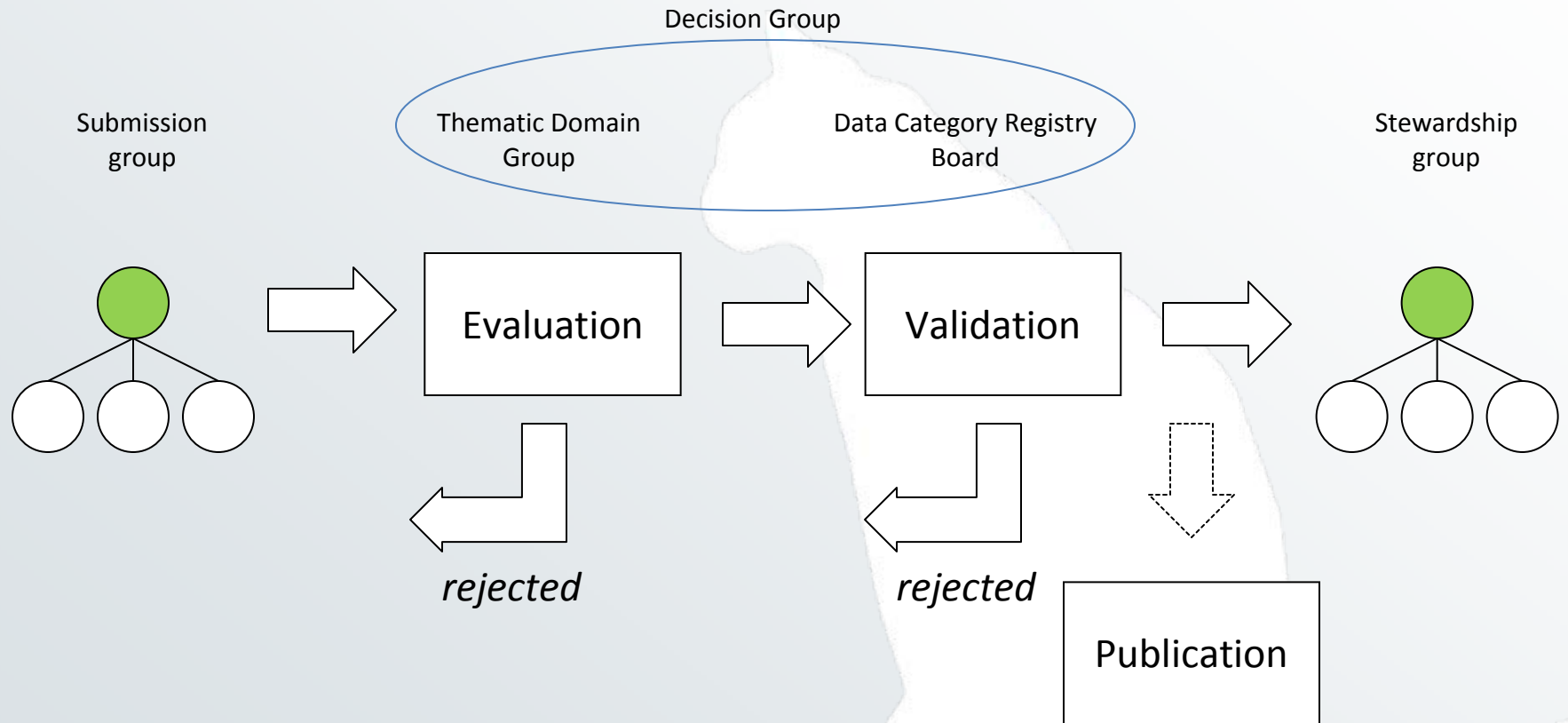
A glimpse of ISOcat



Status of the metadata profile

- Initial set of data categories has been created (to never disappear)
 - Any additional data categories in the pipeline?
 - Your own components might need your own DCs
- Translations for many EU languages have been added
 - Mostly **green** checks
 - Deterioration due to new but incomplete translations
- ISO Standardization
 - TDG ballot is ending this week
 - Implementation of standardization process is ongoing
- The addition of container DCs to be linked to CMD components is planned

Standardization



Metadata Thematic Domain Group (ballot)

- Chair: **Peter Wittenburg (MPI)**
- Members:
 - Thierry Declerck (DFKI)
 - Florian Schiel (Munich)
 - Erhard Hinrichs (Tübingen)
 - Iris Vogel (Tübingen)
 - Claude Martin (LIMIS)
 - Bertrand Gaiffe (ATILF)
 - Maria Gavrilidou (ILSP)
 - Elina Desipri (ILSP)
 - **Daan Broeder (MPI)**
 - **Nelleke Oostdijk (Nijmegen)**
 - Martin Wynne (Oxford)
 - Wim Peters (Oxford)
 - **Helen Aristar-Dry (Michigan)**
 - *Thorsten Trippel (Tübingen)*



Can you trust the DCR?

- Each, also when not standardized, data category has a *Persistent Identifier* (PID)
- The *Registration Authority* of ISO 12620 is obliged to keep these PIDs resolvable
- The DCR is a (core) component of many ISO TC 37 *standards* (TBX, TMF, LMF, LAF, ...)
- The DCR is also a (core) registry in the *CLARIN infrastructure*
- ISOcat is in *beta*, i.e., still actively being developed and not yet feature complete, however, core functionality is *stable* for everyday usage and being backed up every day
 - We welcome any feedback! Contact us: isocat@mpi.nl

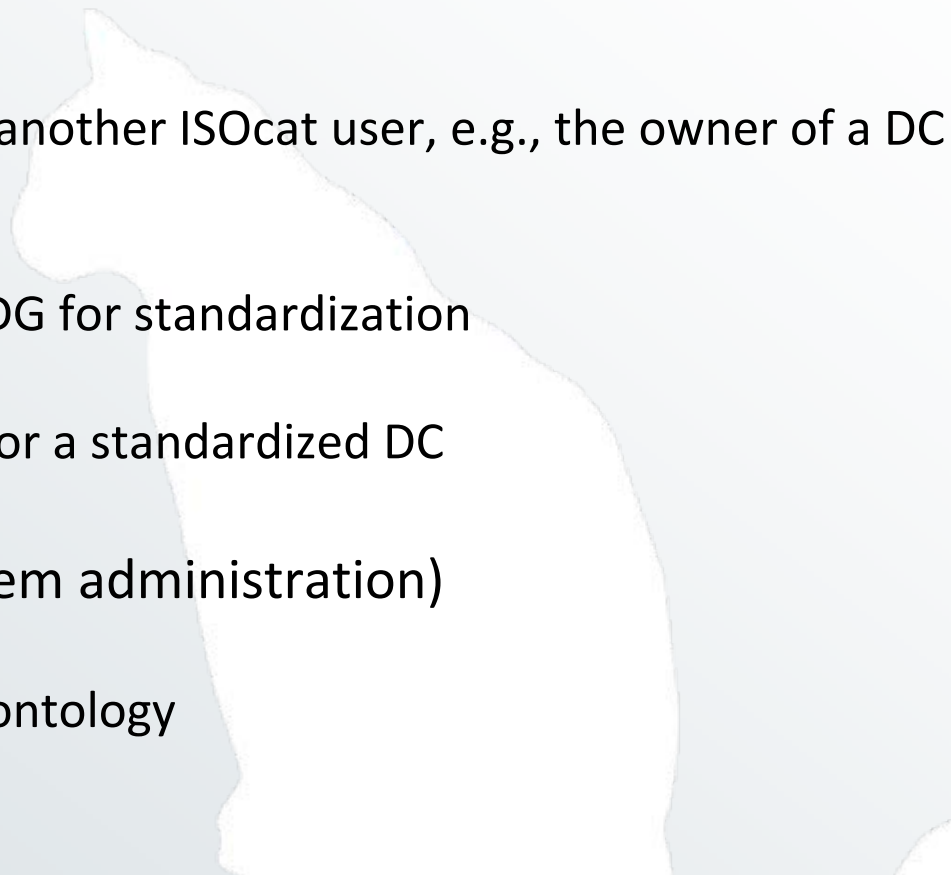
Can you trust a data category?

- Some TDGs are actively *cleaning up* their legacy:
 - Metadata, Morphosyntax and Terminology
- Also at least 2 *new* TDGs are being established
 - Translation and Sign language
- Others are still *asleep* but will hopefully wakeup after the ending of the TDG ballot
 - the DCR development team might interact with them as closely as we do with the current active TDGs to speed them up
- This means DCs are a bit *in flux*
 - Get in touch with the (private) owners (mediated mail is coming soon)
 - Once standardized changes need to go through the standardization process, but old versions will always be available

Can you trust a private data category?

- There are no standardized DCs in the registry yet
 - ISO standardization of a DC is an *optional* step
- Standardized DCs have *extra safe guards*, but might be slow in adapting to changes in the environment
- ISOcat uses a *grass roots* approach
 - Work together in (ad hoc) groups
 - Interact with owners of DCs (email, forum)
 - Become members of TDGs
- There might be some *versioning* policy for privately owned public DCs
 - Freeze the English definition, i.e., a change of the semantics requires a new version (and the old version remains available)

Upcoming

- ISOcat forum
 - A forum per TDG/profile
 - Send a mediated email to another ISOcat user, e.g., the owner of a DC or the chair of a TDG
 - *Standardization support*
 - Submit a set of DCs to a TDG for standardization
 - Standardization workflow
 - Submit a change request for a standardized DC
 - Container DCs
 - DCIF import (by ISOcat system administration)
 - ISO 639 language codes
 - concepts from the GOLD ontology
 - ...
 - Relation Registry
- 

Thank you for your attention!

Visit

www.isocat.org

Questions?

isocat@mpi.nl

